Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Q1 Simultaneous confidence intervals for comparisons of several multinomial samples☆

Q2 Frank Schaarschmidt[a,*], Daniel Gerhard[b], Charlotte Vogel[a]

[a] Leibniz Universitaet Hannover, Institute of Biostatistics, Herrenhaeuserstr. 2, 30419, Hannover, Germany

[b] University of Canterbury, School of Mathematics & Statistics, Private Bag 4800, Christchurch 8041, New Zealand

ARTICLE INFO

Article history:
Received 15 January 2016
Received in revised form 1 September 2016
Accepted 6 September 2016
Available online xxxx

Keywords:
Multiple comparisons
Polytomous data
Dirichlet
Baseline logit
Coverage probability

ABSTRACT

Multinomial data occur if the major outcome of an experiment is the classification of experimental units into more than two mutually exclusive categories. In experiments with several treatment groups, one may then be interested in multiple comparisons between the treatments w.r.t several definitions of odds between the multinomial proportions. Asymptotic methods are described for constructing simultaneous confidence intervals for this inferential problem. Further, alternative methods based on sampling from Dirichlet posterior distributions with vague Dirichlet priors are described. Monte Carlo simulations are performed to compare these methods w.r.t. their frequentist simultaneous coverage probabilities for a wide range of sample sizes and multinomial proportions: The methods have comparable properties for large samples and no rare events involved. In small sample situations or when rare events are involved in the sense that the expected values in some cells of the contingency table are as low as 5 or 10, the method based on sampling from the Dirichlet posterior yields simultaneous coverage probabilities closest to the nominal confidence level. The methods are provided in an R-package and their application is illustrated for examples from developmental toxicology and differential blood counts.

1. Introduction

In a number of toxicological assays, the major outcome is the classification of each experimental unit into one of several categories. For example cells may be classified by visual assessment into several categories, where categories distinguish undamaged cells from different types of unusual characteristics or malformation. In clinical trials, the primary outcome may be the classification of individual patients into one of several categories reflecting disease severity, or clinical subtypes of a certain disease. Often, such categories are ordinal. In some applications, however, the order of categories can be ambiguous, that is, there is no clear order of severity among categories, or there may be no order at all, such that the categories are best described as a nominal variable.

In such trials, multiple treatments can be of interest, for example, multiple dose groups compared to a control group in toxicological assays or different therapeutic interventions in a clinical trial. Counting the number of individuals in each category and each treatment group gives rise to a 2-dimensional contingency table with several rows and columns. In the following, we will assume that the individual experimental units are assigned to treatment groups in a completely

☆ R-code to reproduce the examples and tables containing the simulation settings are available as supplementary material (see Appendix A).

* Correspondence to: Institut fuer Biostatistik, Herrenhaeuser Strasse 2, D-30419 Hannover, Germany. Fax: +49 511 762 4966.
E-mail address: schaarschmidt@biostat.uni-hannover.de (F. Schaarschmidt).

http://dx.doi.org/10.1016/j.csda.2016.09.004
0167-9473/

randomized design and that the sample size per treatment group is fixed by the experimental design (i.e., it is not the result of a random process as, for example, in an epidemiological exposure study). Under these conditions, we may assume that the counts of the different categories follow a multinomial distribution, independently in each treatment group.

Such contingency tables may be analyzed by applying the χ^2 tests for independence. A significant result of such a test will only produce the rather general interpretation: The probability to fall into some of the categories does significantly differ between some of the treatment groups. In practice, this will rarely be an exhaustive interpretation of the data. On the contrary, interest will be in a more detailed interpretation: Which categories increase or decrease in probability between which of the treatment groups, and if so, by what extent? If multiple comparisons between treatments with respect to several categories contribute to an overall hypothesis in the sense of a union intersection test (e.g. Casella and Berger, 2002), simultaneous confidence intervals are necessary for such interpretations. But, depending on the application, not all possible comparisons between categories are of interest and not all comparisons between treatments may play a role for the overall hypothesis. Rather, particular categories and treatments in a given assay or trial will give rise to a special set of comparisons which are of interest.

Methods for simultaneous confidence intervals (SCI) in multiple comparisons in contingency tables have been proposed by Gold (1963) and Goodman (1964). Gold (1963) describes an asymptotic Scheffe-type-approach for SCI suitable for all possible linear combinations of the proportions of several multinomial vectors by using a χ^2-quantile with degrees of freedom as in the corresponding global test. Such approaches are inherently two-sided, and the resulting intervals will be unnecessarily large if only a small subset of comparisons (out of all possible comparisons) is of interest. Goodman (1964) considers asymptotic methods for all possible comparisons as well as a selected subset of comparisons of multinomial proportions on the log-scale, assuming a single multinomial distribution for a contingency table with multiple rows and columns (as suitable, e.g. for epidemiological studies). He shows that Bonferroni-adjusted standard normal quantiles may yield narrower intervals than the Scheffe-type approach, when only few comparisons are of interest. Still this approach can be improved because the Bonferroni-adjustment ignores the correlation between the estimators (or the related test statistics).

Since then, numerous authors have considered simultaneous confidence intervals for proportions or pairwise comparisons of proportions in a single multinomial sample (e.g. Glaz and Sison, 1999; Piegorsch and Richwine, 2001; Hou et al., 2003; Wang, 2000; Chafai and Concordet, 2009). To our knowledge, simultaneous confidence intervals for the comparison of multiple odds between multiple multinomial samples have not been considered any further, although there is room for improvement compared to the seminal methods of Gold (1963) and Goodman (1964): The test statistics related to comparisons of multiple logits of multinomial proportions asymptotically follow a multivariate normal distribution (e.g., Agresti, 2013) and multiple multinomial samples can be considered as a special case for the application of multivariate generalized linear models (e.g. McCullagh and Nelder, 1989; Agresti, 2013). One can thus use quantiles of the multivariate normal distribution (Bretz et al., 2001) based on a sample estimate of the correlation structure to construct asymptotic simultaneous confidence intervals according to Hothorn et al. (2008). Such intervals will be narrower than Bonferroni-adjusted intervals in cases where only a limited subset of parameters with correlated estimators is of interest, because their quantiles account for the correlation structure that is ignored by Bonferroni or Scheffe-type approaches. Although all necessary computational methods are available, these methods have so far not been investigated with respect to their properties when applied with small sample sizes. Also, they suffer from infinite interval bounds, when single cells of the contingency table happen to contain zeros. Further improvements compared to these asymptotic methods might be achievable by sampling from the joint distribution of interest, for example from the posterior of a Bayesian model with a vague prior. Simultaneous confidence intervals can then be computed from such samples by percentile methods as described in Besag et al. (1995), or Mandel and Betensky (2008).

In the remaining part of the paper, we will first describe asymptotic simultaneous confidence intervals for user-defined sets of logits compared between several multinomial samples. Additionally, we will consider simultaneous percentile intervals applied on samples of Dirichlet posteriors with vague Dirichlet priors. The small sample performance of these methods will be compared in frequentist simulation studies. Finally, the methods are applied to two data sets.

2. Material and methods

2.1. Data structure and notation

We consider $g = 1, \ldots, G$ treatment groups in a randomized design, where n_g is the sample size in group g that has been fixed by the experimental design. As the experimental outcome, each individual or experimental unit in group g is categorized into exactly one of C possible categories, with index $c = 1, \ldots, C$. Furthermore we assume that due to the randomized assignment of treatments to individuals or experimental units, there is no further subgrouping of individuals or heterogeneity among individuals and also, that there are no secondary factors or covariates that affect the outcome. Thus we assume that the counted number of individuals of categories $c = 1, \ldots, C$ in group g, $\mathbf{x}_g = (x_{g1}, x_{g2}, \ldots, x_{gC})$, follows a multinomial distribution

$$(x_{g1}, x_{g2}, \ldots, x_{gC}) \sim multinomial\left(n_g, (\pi_{g1}, \pi_{g2}, \ldots, \pi_{gC})\right),$$

Please cite this article in press as: Schaarschmidt, F., et al., Simultaneous confidence intervals for comparisons of several multinomial samples. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.09.004

where π_{gc} is the unknown probability of an individual in treatment group g to fall into category c. Usually, such observed counts are summarized in a contingency table, $\mathbf{X}_{(G\times C)}$.

2.2. Parameters of interest

A simple choice for the analysis of such data is to compare the baseline logits between the groups. That is, the ratios of the latter proportions, $\pi_{g2}, \ldots, \pi_{gC}$, to that of the first category π_{g1} (the baseline category) are of interest. Treatment effects are then expressed as the relative change of these ratios between the treatment groups. For only two treatment groups, $g = 1, 2$, the odds ratios of interest are then

$$\left(\frac{\pi_{22}/\pi_{21}}{\pi_{12}/\pi_{11}}, \frac{\pi_{23}/\pi_{21}}{\pi_{13}/\pi_{11}}, \ldots, \frac{\pi_{2C}/\pi_{21}}{\pi_{1C}/\pi_{11}}\right).$$

Depending on the practical meaning of the different categories, more or less parameters than these comparisons to the baseline categories might be of interest. Either, the comparisons of certain categories to baseline may be not of primary interest, or, additional odds, referring to ratios between the proportions of categories $c = 2, \ldots, C$, may be important. On the log scale, all possible pairwise logits can be written as

$$\begin{pmatrix}\delta_{g1}\\ \delta_{g2}\\ \vdots\\ \delta_{gI}\end{pmatrix} = \mathbf{A}_{(I\times C)} \log\left(\boldsymbol{\pi}_g^T\right) = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0\\ -1 & 0 & 1 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & & \vdots & \vdots\\ -1 & 0 & 0 & \cdots & 0 & 1\\ 0 & -1 & 1 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \begin{pmatrix}\log(\pi_{g1})\\ \log(\pi_{g2})\\ \vdots\\ \log(\pi_{gC})\end{pmatrix}.$$

Note that on the scale of baseline logits, $\psi_{gc} = \log(\pi_{gc}) - \log(\pi_{g1})$, $c = 2, \ldots, C$, all pairwise logits can be written as

$$\begin{pmatrix}\delta_{g1}\\ \delta_{g2}\\ \vdots\\ \delta_{gI}\end{pmatrix} = \mathbf{A}^*_{(I\times C-1)} \boldsymbol{\psi}_g^T = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0\\ 0 & 1 & \cdots & 0 & 0\\ \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & \cdots & 0 & 1\\ -1 & 1 & \cdots & 0 & 0\\ \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \begin{pmatrix}\psi_{g2}\\ \psi_{g3}\\ \vdots\\ \psi_{gC}\end{pmatrix}.$$

2.3. Between-group comparisons of interest

Similarly, comparisons to a control treatment ('Dunnett-type'), all pairwise comparisons ('Tukey-type') between treatments or a particular subset of these may be of interest, depending on the practical meaning of the G treatment groups for a given experimental question. We can thus write the between-group-comparisons in a contrast matrix $\mathbf{B}_{(J\times G)}$ for the ith logit defined above

$$\boldsymbol{\theta}_i = \begin{pmatrix}\theta_{1i}\\ \theta_{2i}\\ \vdots\\ \theta_{Ji}\end{pmatrix} = \mathbf{B}_{(J\times G)} \boldsymbol{\delta}_i^T = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0\\ -1 & 0 & 1 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & & \vdots & \vdots\\ -1 & 0 & 0 & \cdots & 0 & 1\\ 0 & -1 & 1 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \begin{pmatrix}\delta_{1i}\\ \delta_{2i}\\ \vdots\\ \delta_{Gi}\end{pmatrix}.$$

If these between-group-comparisons $j = 1, \ldots, J$ are the same for all logits $i = 1, \ldots, I$, the parameter vector can be briefly written as

$$\boldsymbol{\theta} = (\mathbf{B} \otimes \mathbf{A}) \begin{pmatrix}\log \boldsymbol{\pi}_1\\ \log \boldsymbol{\pi}_2\\ \vdots\\ \log \boldsymbol{\pi}_G\end{pmatrix},$$

where the elements θ_{ij} of $\boldsymbol{\theta}$ are primarily ordered by between group comparison $j = 1, \ldots, J$ and then, for each j, by odds ratio $i = 1, \ldots, I$ (inner order).

Please cite this article in press as: Schaarschmidt, F., et al., Simultaneous confidence intervals for comparisons of several multinomial samples. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.09.004

2.4. Simultaneous Wald-type confidence intervals

The statistical model outlined above is a special case of a multivariate generalized linear model (Agresti, 2013; McCullagh and Nelder, 1989), for which the baseline logits $\psi_{gc}, c = 2, \ldots, C$ are the natural parameter (Agresti, 2013). One can thus use the methods of Hothorn et al. (2008) to construct simultaneous confidence intervals for $\boldsymbol{\theta}$ based on the estimated baseline logits $\hat{\boldsymbol{\psi}}$ and the corresponding estimated variance covariance matrix $\hat{\boldsymbol{\Sigma}}$. In more general settings, such estimates could be obtained by fitting baseline logit models. Then, also secondary factors or covariates might be included in such a model. In the simple case considered here, these estimates can be obtained from the contingency table $\mathbf{X}_{(G\times C)}$, using the asymptotic variance of baseline logits under multinomial sampling (derived using the Delta Method in Agresti, 2013, p.590–591): Denote the vector of sample estimators of the log-proportions in group g by $\log\left(\hat{\boldsymbol{\pi}}_g\right) = \left(\log\left(x_{g1}/n_g\right), \ldots, \log\left(x_{gC}/n_g\right)\right)$. The corresponding estimators of the I linear combinations of interest are $\hat{\boldsymbol{\delta}}_g = \mathbf{A}\log\left(\hat{\boldsymbol{\pi}}_g\right)$, which have the asymptotic covariance matrix (Agresti, 2013, p. 591).

$$\boldsymbol{\Sigma}_g = n_g^{-1}\left(\mathbf{A}\mathtt{Diag}(\boldsymbol{\pi}_g)^{-1}\mathbf{A}^T - \mathbf{A}\mathbf{1}\mathbf{1}^T\mathbf{A}^T\right),$$

where $\mathtt{Diag}(\boldsymbol{\pi}_g)^{-1}$ is the inverse of a diagonal matrix containing the true proportions $\boldsymbol{\pi}_g$, and $\mathbf{1}$ is an $(C \times 1)$ vector with all elements = 1.

Since we assumed independence between the treatment groups $g = 1, \ldots, G$, we can assemble the logits of interest for all treatment groups $g = 1, \ldots, G$ by stacking the column vectors $\boldsymbol{\delta}_g$, such that the corresponding covariance matrix can be written as a block-diagonal matrix:

$$\boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \\ \vdots \\ \boldsymbol{\delta}_G \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_G \end{pmatrix}.$$

The between-group comparisons (outer order) for all logits of interest (inner order) can then be written as

$$\boldsymbol{\theta} = \left(\mathbf{B} \otimes \mathbf{I}_{I\times I}\right)\boldsymbol{\delta},$$

where $\boldsymbol{I}_{I\times I}$ is the identity matrix with I rows and columns. The corresponding covariance matrix is

$$\mathbf{V} = \left(\mathbf{B} \otimes \mathbf{I}_I\right)\boldsymbol{\Sigma}\left(\mathbf{B} \otimes \mathbf{I}_I\right)^T.$$

Estimators for $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}_g$, $\boldsymbol{\Sigma}$ and $\mathbf{V}$, may be obtained by plugging-in of the sample proportions $\hat{\boldsymbol{\pi}}_g$ instead of $\boldsymbol{\pi}_g$, and are denoted as $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\Sigma}}_g$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{\mathbf{V}}$. Approximate simultaneous confidence intervals for the $M = IJ$ corresponding odds ratios are then

$$\exp\left[\hat{\theta}_m \pm z_{\mathtt{two-sided},1-\alpha,M,\hat{\mathbf{R}}}\sqrt{\hat{v}_m}\right], \quad m = 1, \ldots, M,$$

where $\hat{\theta}_m$ is the mth element of $\hat{\boldsymbol{\theta}}$, $\hat{v}_m$ is the mth element of diagonal of $\hat{\boldsymbol{\Sigma}}$, $z_{\mathtt{two-sided},1-\alpha,M,\hat{\mathbf{R}}}$ is the two-sided $1-\alpha$-quantile of the M-variate normal distribution (Genz and Bretz, 2009) with correlation matrix $\hat{\mathbf{R}}$, and $\hat{\mathbf{R}}$ is obtained by standardizing $\hat{\boldsymbol{\Sigma}}$ by its diagonal elements (Hothorn et al., 2008).

Clearly, this approach has a number of problems: The plug-in of $\hat{\boldsymbol{\pi}}_g$ to obtain $\hat{\boldsymbol{\Sigma}}$, and the plug-in of $\hat{\boldsymbol{R}}$ to obtain the multivariate normal quantile $z_{\mathtt{two-sided},1-\alpha,M,\hat{\mathbf{R}}}$ are only justified for large samples (Hothorn et al., 2008). Moreover, $\boldsymbol{\Sigma}$ is only the asymptotic variance. The confidence intervals are symmetric with respect to $\hat{\theta}_m$, but the sampling distribution of $\hat{\theta}_m$ may be skewed if some expected cell counts, $n_g\pi_{gc}$, are small and π_{gc} differ, that is, if some sample sizes are moderate and/or the proportions are close to the border of the parameter space. Finally, the plug-in of π_{gc} with extreme observations as $x_{gc} = 0$ yields unreasonable estimated variances (∞) for the parameters on the log-scale; this leads to the failure of computing $z_{\mathtt{two-sided},1-\alpha,M,\hat{\mathbf{R}}}$, based on $\hat{\mathbf{R}}$, and even when using some ad-hoc adjustment for computing $\hat{\mathbf{R}}$, the intervals involving the corresponding π_{gc} will be uninformative due to spanning the complete parameter space. In parameter settings, where such events occur frequently, we can expect that the Wald-type simultaneous confidence intervals are unnecessarily conservative, that is, cover the true parameters too often.

In order to deal with the last problem, we apply the following ad-hoc adjustments: To compute the correlation matrix and contrasts of interest when the contingency table contains zeros, these are replaced by 0.5 (e.g. Plackett, 1962; Goodman, 1964). This approach is referred to as **W**. Alternatively, one may use $\tilde{x}_{gc} = x_{gc} + 0.5$, $\tilde{n}_g = \sum_{c=1}^{C}\tilde{x}_{gc}$ and $\tilde{\boldsymbol{\pi}}_g = \left(\tilde{x}_{g1}/\tilde{n}_g, \tilde{x}_{g2}/\tilde{n}_g, \ldots, \tilde{x}_{gC}/\tilde{n}_g\right)$ instead of $\hat{\boldsymbol{\pi}}_g$ in all computations above. That is, 0.5 is added to each cell of the $G\times C$ contingency table, and all subsequent computations are performed based on this altered contingency table. This adjusted method is referred to as **W0.5**.

2.5. Sampling from the posterior distribution with a weakly informative prior

Under the assumption of G independent multinomial samples, one can make use of the fact that the Dirichlet distribution is a conjugate prior for the assumption of multinomial data. In the Bayesian model

Please cite this article in press as: Schaarschmidt, F., et al., Simultaneous confidence intervals for comparisons of several multinomial samples. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.09.004

$$(\pi_{g1}, \pi_{g2}, \ldots, \pi_{gC}) \sim \textit{Dirichlet}\left((\alpha_{g1}, \alpha_{g2}, \ldots, \alpha_{gC})\right),$$
$$(x_{g1}, x_{g2}, \ldots, x_{gC}) \sim \textit{multinomial}\left(n_g, (\pi_{g1}, \pi_{g2}, \ldots, \pi_{gC})\right),$$

we can easily draw samples from the joint posterior distribution,

$$P\left((\pi_{g1}, \ldots, \pi_{gC}) \mid (x_{g1}, \ldots, x_{gC})\right) \sim \textit{Dirichlet}\left((x_{g1} + \alpha_{g1}, \ldots, x_{gC} + \alpha_{gC})\right).$$

To construct simultaneous intervals for $\boldsymbol{\theta}$, many (say K) samples are drawn from this posterior, independently for each group g: Denote with $\boldsymbol{p}_k$ the stacked vectors of all groups $g = 1, \ldots, G$ in the kth sample, that is, $\boldsymbol{p}_k = (p_{11}, \ldots, p_{1C}, p_{21}, \ldots, p_{2C}, p_{G1}, \ldots, p_{GC})^T$. For each sample $k = 1, \ldots, K$, the corresponding sample for the $M = IJ$ parameters of interest can be computed by:

$$\boldsymbol{t}_k = (\mathbf{B} \otimes \mathbf{A}) \log \boldsymbol{p}_k.$$

Rectangular sets containing the central 95% of the K sampled vectors $\boldsymbol{t}_k$, $k = 1, \ldots, K$ are described by Besag et al. (1995). For a $(K \times M)$ matrix $\boldsymbol{T}$, containing the K samples of the parameter vector of interest, $\boldsymbol{t}_k$, the main steps of this procedure are recalled here in close relation to the descriptions in Schaarschmidt and Djira (in press) or Schaarschmidt (2013):

1. Rank each column, $m = 1, \ldots, M$ of $\boldsymbol{T}$ separately and record the resulting ranks r_{km} and order statistics $t_m^{(k)}$.
2. For each row, $k = 1, \ldots, K$, of the resulting matrix $(K \times M)$ matrix of ranks with elements r_{km}, compute $\max_k = \max(\max_{m=1,\ldots,M}(r_{km}), K + 1 - \min_{m=1,\ldots,M}(r_{km}))$.
3. Order $\max_k$, resulting in the order statistics $\max^{[k]}$ and find $k^* = \max^{[q_{0.95}]}$, where $q_{0.95}$ is the nearest integer to $K * 0.95$.

The lower and upper interval bounds for each parameter of interest, $m = 1, \ldots, M$, are then obtained from the order statistics and back-transformation to the scale of odds-ratios: $\exp\left[t_m^{(K+1-k^*)}, t_m^{(k^*)}\right]$. Corresponding one-sided 95% simultaneous percentile intervals (Mandel and Betensky, 2008) can be calculated to obtain upper and/or lower limits for each element of $\boldsymbol{\theta}$. When the prior is chosen such that it has nearly no impact on the posterior, one can expect that the resulting intervals have good frequentist properties, that is, simultaneous coverage probability close to 95%. Choosing the prior parameters $\alpha_{gc} = 1$ for all g, c corresponds to a uniform prior distribution for $C = 2$, while $\alpha_{gc} = 0.5$ for all g, c is known as Jeffreys prior. In the following, such intervals will be called **DP0.5** and **DP1**.

2.6. Simulation study

In order to compare the frequentist coverage probabilities between the different methods, a Monte Carlo simulation has been performed for the following parameter settings: for $C = 3$ or $C = 5$ categories and $G = 4$ treatment groups, balanced sample sizes per treatment group of $n_g = 10, 20, 50, 100, 1000$ are considered. Three different sets of odds ratios have been considered: Baseline logits are compared between treatments ($g = 2, 3, 4$) and the control group ($g = 1$), as well as all pairwise comparisons between treatment groups for baseline logits, and all pairwise logits compared between treatments and control group. The true proportions of the categories are varied from the case that all categories appear equally often (1/3, 1/3, 1/3) to settings where the earlier categories (serving as baseline) are dominating (up to $\pi_{g1} = 0.9$) and the remaining categories are rare (down to $\pi_{g3} = 0.01$), and conversely, settings where the earlier categories are rare ($\pi_{g1} = 0.01$) and the remaining categories are abundant ($\pi_{g3} = 0.9$). For $C = 3$, 59 different configurations of $\boldsymbol{\pi}_g$ have been simulated. In 21 of these, all logits are equal between all treatment groups, in the remaining 38 settings some logits differ between some of the treatment groups. For $C = 5$, 35 different parameter settings for $\boldsymbol{\pi}_g$ were considered (13 implying equal logits between treatment groups and 22 implying differences); with five categories, only samples sizes $n_g = 50, 100, 1000$ per group have been considered. A complete list of parameter settings is available as supplementary material (see Appendix A). For each resulting parameter setting, 1000 data sets have been drawn from the multinomial distribution. The methods based on sampling from the Dirichlet distribution have been applied with $K = 10{,}000$ values drawn from the posterior to compute the simultaneous intervals for each data set.

2.7. Software

An implementation of the Wald-type intervals is available in the R-package `MCPAN 1.1-20` (Schaarschmidt et al., 2016) relying on multivariate normal quantiles obtained from the R-package `mvtnorm` (Genz et al., 2015). The methods based on sampling from the Dirichlet-posterior make use of the R-package `MCMCpack` (Martin et al., 2011) for Dirichlet random numbers, and the percentile intervals (Besag et al., 1995; Mandel and Betensky, 2008) implemented in package `MCPAN`.

3. Results

3.1. Simultaneous coverage probabilities

Fig. 1 shows the simulated simultaneous coverage probabilities (SCP) of nominal 95% simultaneous confidence intervals. For all methods, there is a clear dependency of the SCP on the minimal expected cell count ($\min(n_{gc}\pi_{gc})$): Intervals cover

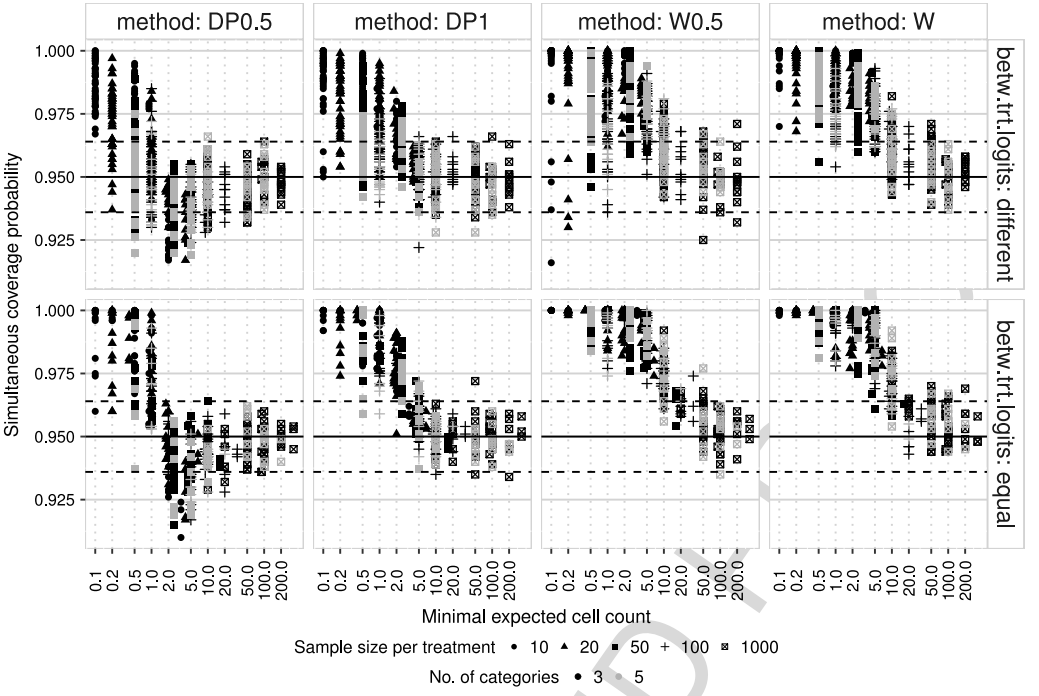

Fig. 1. Simultaneous coverage probabilities of nominal 95% simultaneous confidence intervals, in dependence of the minimal expected cell count $\min(n_{gc}\pi_{gc})$. Symbols distinguish sample size per treatment group g, grayscale distinguishes parameter settings with $C = 3$ or $C = 5$ categories. Column panels show results for different confidence intervals methods, while row panels distinguish parameter settings where at least one logit differs between treatment groups (upper row) and all logits of interest are equal in all treatment groups (lower row). Dashed horizontal lines show the range in which 95% of all simulation results (based on 1000 simulations per setting) would fall if a method had exactly 95% true simultaneous coverage probability.

the parameters too often if the minimal expected cell count is small, that is below 5 or 2, while SCPs are close to the nominal level when the minimal expected cell count is equal or larger than 50. The intervals based on sampling from the Dirichlet posterior with uniform priors (DP1) show SCPs close to or above the nominal levels, whereas using Jeffreys prior may result in SCPs below the nominal level. The DP1 interval shows improved SCP compared to the Wald-type interval for intermediate values of the minimal expected cell count: While the Wald-type intervals start to be too conservative for minimal expected cell counts in the range of 10 or 20, the DP1 method shows SCPs close to the nominal level for minimal expected cell counts of 5 or 10. With the exception of a few parameter settings, the ad-hoc approach of adding 0.5 to each cell and using the Wald-type intervals afterwards does not show tangible differences of the SCP. Fig. 2 illustrates the improvement of SCP when using the DP1 method instead of the Wald-type interval: With sample sizes such as 100, 50, or 20 per group, the DP1 is less conservative than the Wald-type interval for the majority of parameter settings but rarely shows observed SCP larger than that of the Wald-type interval.

As a secondary criterion, we consider the potential imbalance of the lower and upper limits of marginal intervals with respect to the probability to exclude the true parameter. Fig. 3 shows the difference of probabilities to exclude the true parameter between lower and upper limits, scaled by the total for the corresponding parameter setting. Values close to 0 indicate that the probabilities to exclude the true parameter are balanced between the upper and lower limits of marginal intervals, while values approaching -1 or 1 indicate that the interval is biased w.r.t. to the probability to exclude the true parameter. While the W and W0.5 method show similar amounts of unbalanced tail probabilities for given parameter settings, the DP1 method shows reduced imbalance compared to the W method for many parameter settings.

4. Examples

4.1. Developmental toxicity

Agresti (1990, p.320, Tab. 9.7 therein) shows results of a study on developmental toxicity in mice. After exposure to $G = 5$ treatments (control d0, and 4 different dosages, d62.5, d125, d150, d500) of a compound during pregnancy, the offspring of the mice ($n_1 = 297$, $n_2 = 242$, $n_3 = 312$, $n_4 = 299$, $n_5 = 285$) is classified into $C = 3$ categories: alive, dead, malformation. Fig. 4 shows a mosaic plot derived from the (5×3) contingency table data. To investigate for which dose groups there is an increase of π_{dead}/π_{alive} or $\pi_{malformation}/\pi_{alive}$ over that of the control, one can consider simultaneous confidence intervals for baseline logits (baseline = alive) compared between the 4 dose groups and the control.

Please cite this article in press as: Schaarschmidt, F., et al., Simultaneous confidence intervals for comparisons of several multinomial samples. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.09.004

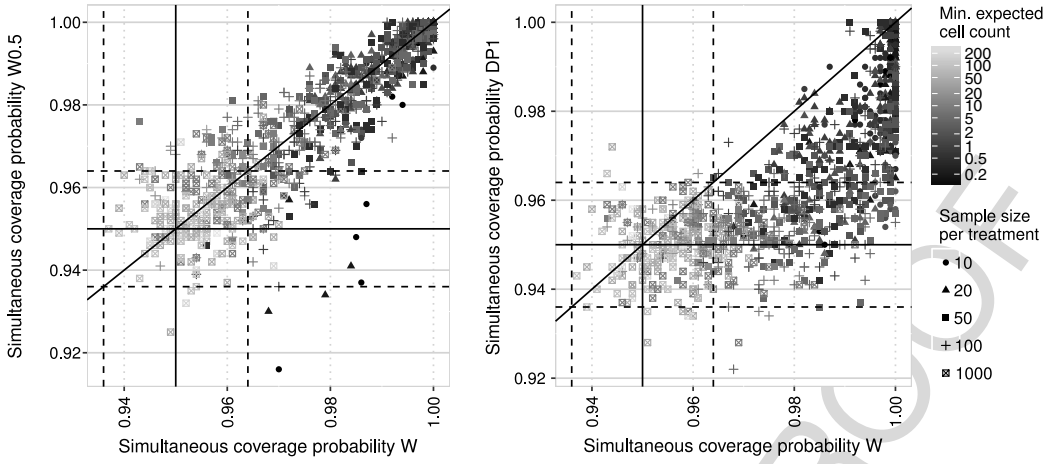

Fig. 2. Simulated (1000 simulation runs per parameter setting) simultaneous coverage probabilities of the Wald-type interval (x-axis) plotted against that of the Wald-add-0.5 interval and the interval based on Dirichlet sampling (DP1). Gray scale is used to show each settings minimal expected cell count $\min(n_{gc}\pi_{gc})$; symbols distinguish sample size per treatment group g. Dashed horizontal and vertical lines show the range in which 95% of all simulation results (1000 simulations) would fall if a method had exactly 95% true coverage probability.

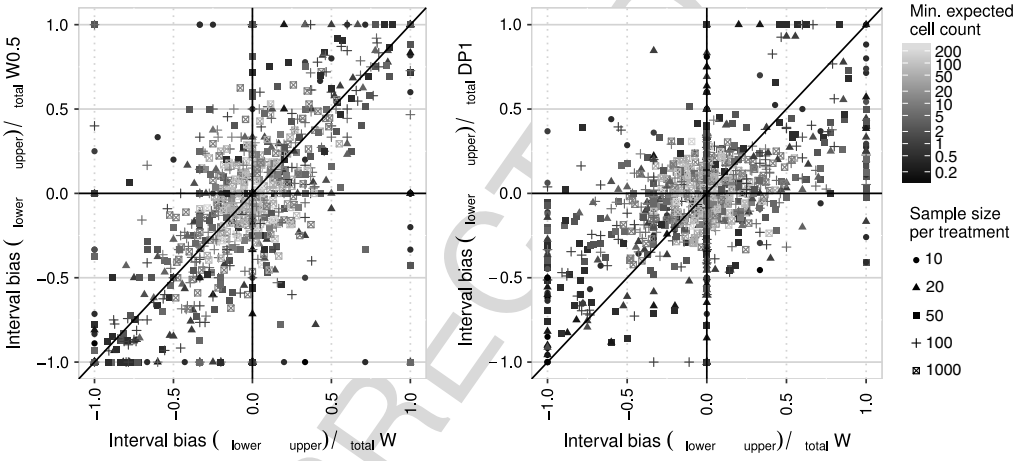

Fig. 3. Difference of probabilities to exclude the true parameter by lower and upper limits of the intervals (1000 simulation runs per parameter setting). Results for the Wald-type interval (x-axis) are plotted against those of the Wald-add-0.5 interval and the interval based on Dirichlet sampling (DP1). Gray scale is used to show each settings minimal expected cell count $\min(n_{gc}\pi_{gc})$; symbols distinguish sample size per treatment group g.

Fig. 5 shows a scatter plot matrix of 2000 sampled values for the ($M = 8$) corresponding logits based on a sample of the joint posterior with prior $(\pi_{alive}, \pi_{dead}, \pi_{malformation}) \sim Dirichlet((1, 1, 1))$ on each sample $g = 1, \ldots, 5$. It is obvious that those parameters referring to comparisons to the control group for the same odds are positively correlated (parameters $1, \ldots, 4$ and $5, \ldots, 8$, respectively) and that the magnitude of correlation further depends on the estimated proportions (higher positive correlations in malformed/alive than in dead/alive). Fig. 6 shows the estimated correlation matrix ($\hat{\boldsymbol{R}}$) underlying the quantiles Wald-type-intervals (W) for this example. Table 1 shows the lower and upper limits of the corresponding 95% simultaneous intervals for the oddsratios: The odds dead/alive are significantly increased compared to control in d250 and d500. According to the DP1 method this ratio is increased by factor 1.5–8.3 in d250 and by factor 97–930 in d500. The odds malformation/alive also show a significant increase in dose groups d250 and d500, at least by factor 10 and 390 (DP1), respectively. The R code to reproduce these calculations (up to uncertainties due to sampling) is provided as supplementary material (see Appendix A).

The corresponding (two-sided) 95% quantile of an 8-variate normal distribution is $z_{0.95,M=8,\hat{\boldsymbol{R}}} = 2.638$. Compared to the Scheffe-type quantile adjusting for all possible contrasts (Gold, 1963), $\sqrt{\chi^2_{df=8}} = 3.938$, the Wald-type intervals have considerably reduced width, whereas the reduction of width is relatively little compared to the Bonferroni adjustment of Goodman (1964): $z_{1-0.05/(8*2)} = 2.734$. Q3

Please cite this article in press as: Schaarschmidt, F., et al., Simultaneous confidence intervals for comparisons of several multinomial samples. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.09.004

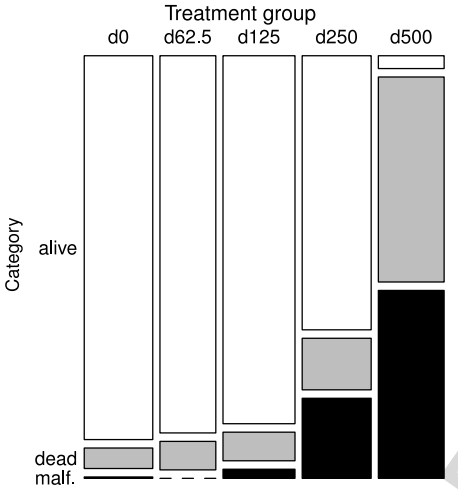

Fig. 4. Mosaicplot of the (5 × 3) table of developmental toxicity data (Agresti, 1990).

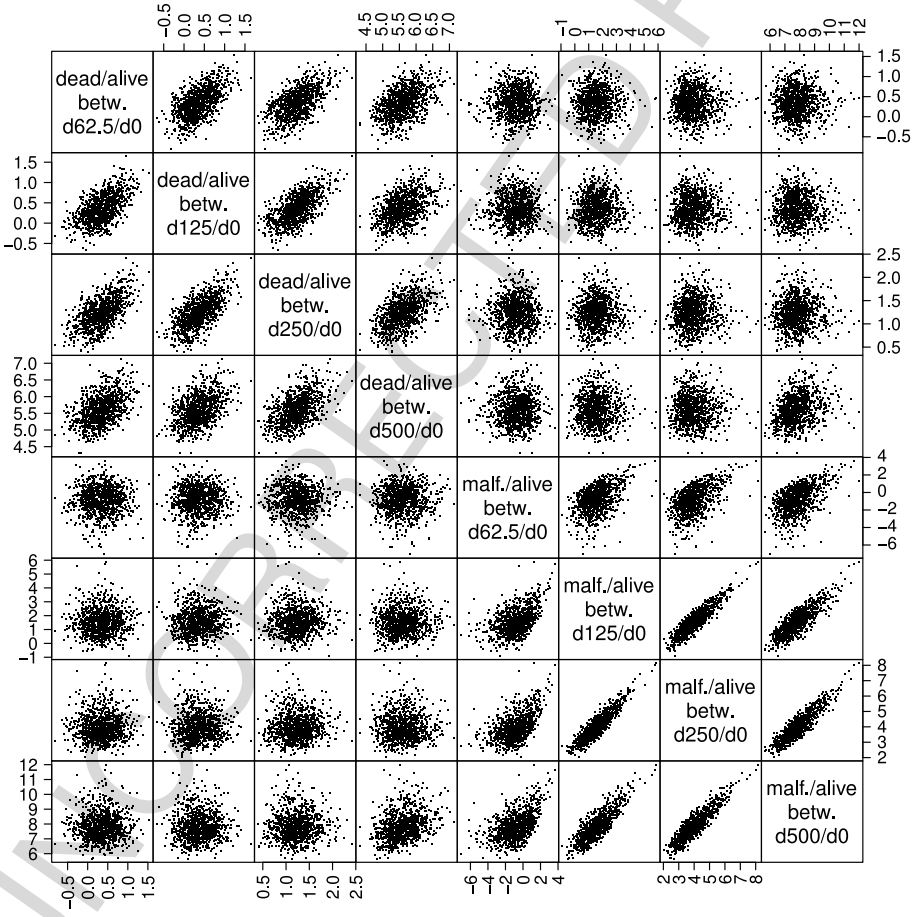

Fig. 5. Sample of 1000 values from the Dirichlet posterior with uniform prior (DP1).

4.2. Differential blood count (WBC) in rats

Table 2 (Hothorn et al., 2009) shows counts of white blood cells of 4 categories, LY, MO, NE, EO (lymphocytes, monocytes, neutrophil and eosinophil granulocytes); other cell types occurred only with one cell and are omitted. Counts have been obtained from rats (females and males) under four different treatments: an untreated control (C) and three dose groups

Please cite this article in press as: Schaarschmidt, F., et al., Simultaneous confidence intervals for comparisons of several multinomial samples. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.09.004

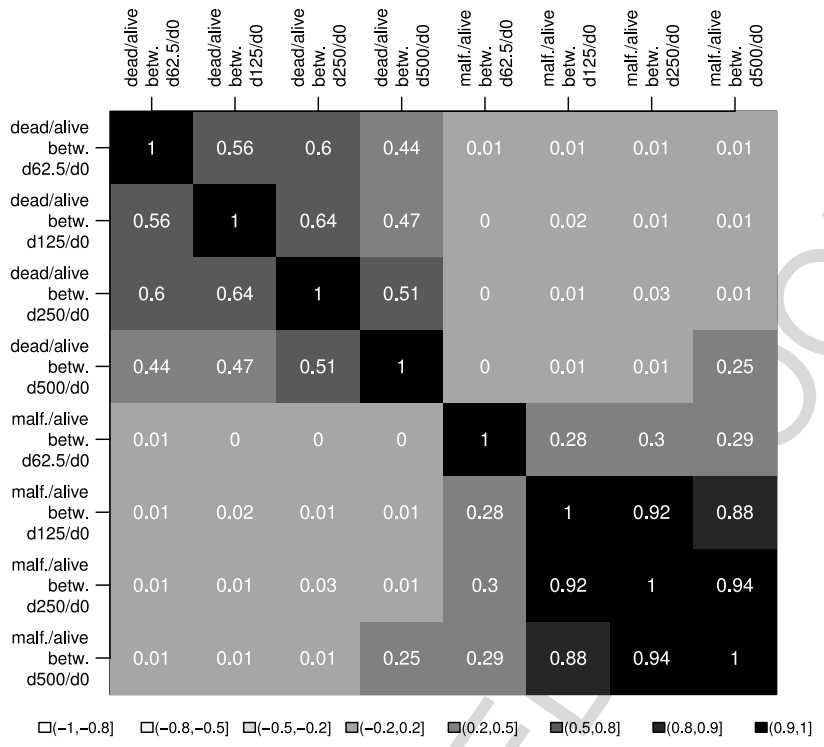

Fig. 6. Estimated correlation matrix $\hat{\boldsymbol{R}}$ corresponding underlying the Wald-type interval (W).

Table 1
Simultaneous 95% confidence intervals for comparisons to control for the baseline odds dead/alive and malformed/alive, rounded to the second significant digit.

Oddsratio	Comparison	Estimate	W		DP1	
			Lower	Upper	Lower	Upper
dead/alive	d62.5/d0	1.4	0.54	3.7	0.55	3.7
dead/alive	d125/d0	1.5	0.59	3.6	0.59	3.7
dead/alive	d250/d0	3.5	1.5	8.2	1.5	8.3
dead/alive	d500/d0	300	95	940	97	930
malf./alive	d62.5/d0	0.62	0.01	60	0.00	18
malf./alive	d125/d0	6.9	0.41	120	0.66	88
malf./alive	d250/d0	82	5.7	1200	10	890
malf./alive	d500/d0	4100	250	67 000	390	45 000

Table 2
Differential count of white blood cells in rats of both sexes and four treatment groups.

Sex	Group	LY	MO	NE	EO
Female	C	1668	41	272	19
Female	L	1633	47	305	15
Female	M	1699	39	244	18
Female	H	1643	37	299	21
Male	C	1594	32	340	34
Male	L	1593	25	356	26
Male	M	1510	34	431	25
Male	H	1196	33	351	19

(L, M, H, for low, mid and high dose). Note that the counts in Table 2 are obtained by pooling ten individuals per sex and treatment group (exception: eight animals for males in high dose).

One may now be interested, whether any of the relative proportions of the single categories changes between the control and the dose groups in males or females. We express this as all ($I = 6$) pairwise odds between the $C = 4$ categories. These odds are then compared between the L, M and H dose and the control, separately for males and females, resulting in $J = 6$

Please cite this article in press as: Schaarschmidt, F., et al., Simultaneous confidence intervals for comparisons of several multinomial samples. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.09.004

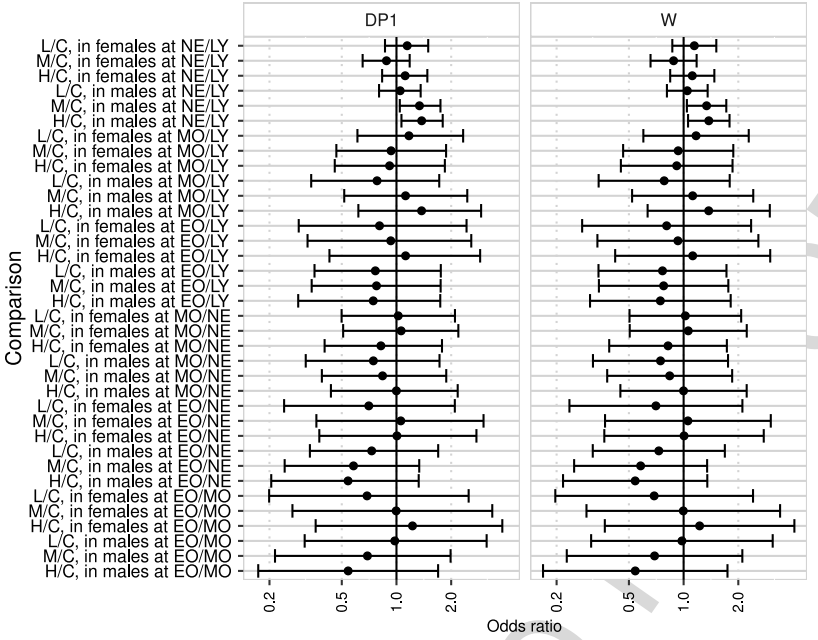

Fig. 7. Simultaneous 95% confidence intervals for 36 odds ratios defined in the differential blood count example.

Table 3
Simultaneous 95% confidence intervals for comparisons to control for the baseline odds dead/alive and malformed/alive (subset out of a total of 36 comparisons), rounded to the 3rd digit.

Oddsratio	Estimate	W		DP1	
		Lower	Upper	Lower	Upper
NE/LY btw L/C, fem.	1.145	0.868	1.511	0.868	1.511
NE/LY btw M/C, fem.	0.881	0.659	1.178	0.658	1.183
NE/LY btw H/C, fem.	1.116	0.845	1.474	0.844	1.472
NE/LY btw L/C, mal.	1.048	0.810	1.355	0.810	1.357
NE/LY btw M/C, mal.	1.338	1.044	1.716	1.041	1.720
NE/LY btw H/C, mal.	1.376	1.059	1.787	1.065	1.792

between-group-comparisons. The corresponding matrices A and B are then:

$$\mathbf{A}_{(I\times C)} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{B}_{(J\times G)} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix}.$$

The counts in Table 2 are relatively large, thus all considered methods can be expected to perform well and to yield very similar results. The quantiles of the Goodman approach ($z_{1-0.05/(36*2)} = 3.197$) and the Wald-type intervals with plug-in of estimated correlations ($z_{0.95,M=36,\hat{\boldsymbol{R}}} = 3.085$) again differ only slightly.

Fig. 7 shows the 95% confidence intervals for the 36 odds ratios defined above for the methods DP1 and W. The intervals hardly differ between the methods. Significant differences with respect to these 36 odds ratios are found for the mid and high dose groups (H, M) in males where the proportion of neutrophil granulocytes relative to lymphocytes (π_{NE}/π_{LY}) is significantly increased in treatment groups M and H compared to the control group, C. Table 3 shows estimates and confidence limits of the W and DP1 method for those odds (π_{NE}/π_{LY}): in males, the ratio (π_{NE}/π_{LY}) is increased by factor [1.041, 1.720] in group M, and by factor [1.065, 1.792] in group H, relative to that of the control group. The R code to reproduce these calculations is provided as supplementary material (see Appendix A).

5. Discussion

We described methods for the computation of simultaneous confidence intervals for user defined sets of pairwise between-treatment comparisons and user-defined sets of odds ratios based on the assumption of several independent multinomial samples. The asymptotic method accounts for the correlation between estimators by the plug-in of an estimated

Please cite this article in press as: Schaarschmidt, F., et al., Simultaneous confidence intervals for comparisons of several multinomial samples. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.09.004

covariance matrix. A small sample approach, based on sampling from a Dirichlet posterior with vague priors, is considered as an alternative.

In the simulation study, the coverage probability of these methods is assessed for different sets of multinomial proportions, different sample sizes per treatment group, three (or five) multinomial categories, as well as different types of comparisons between groups and categories. The method based on sampling from the Dirichlet posterior with a vague prior assigning parameter $\alpha = 1$ to all categories performs best in the considered settings: When the minimal expected cell count of the contingency table is moderate (e.g. at least five) the simultaneous coverage probability is close to the nominal level. If rare proportions or small sample sizes lead to smaller expected cell counts, the method is conservative. The asymptotic method is more conservative as it shows coverage probabilities close to the nominal level for expected cell counts of 20 or above, and covers the true parameter too often otherwise. Note that these recommendations may not hold when comparing multinomial samples with much more categories than considered here, e.g. 10 or 20.

All methods considered are conservative for small sample size and/or rare events. That is, with either method it will be hard to detect relatively small changes in the proportions of rare categories, or when sample sizes are small. The method based on sampling from the Dirichlet posterior can easily be extended to include informative priors. For example, when historical control data are available for bioassays, the Dirichlet prior for untreated control groups may be chosen to reflect the expected values and the plausible range for the proportions of the categories under control conditions. Moreover, it would be computationally simple, to extend the methods based on Dirichlet posteriors to simultaneous confidence intervals for differences or ratios between multinomial proportions.

One may still want to use exact methods to construct intervals for comparing the proportions of several categories between several groups. A straightforward way to use exact methods could be based on the fact that the binomial distribution is the marginal distribution of the multinomial: when considering a given category c as success and all remaining categories as failures, various methods to compute exact confidence intervals are available for the comparison of two binomial samples (e.g., Chan and Zhang, 1999; Reiczigel et al., 2008; Wang and Shan, 2015). However, this would not offer inference for that set of parameters that has been considered above. A Bonferroni-correction of individual intervals would be needed to obtain simultaneous confidence intervals. Assuming one multinomial sample, Hayter (2014) describes the efficient computation for multinomial probabilities, and Chafai and Concordet (2009) and Wang (2000) consider simultaneous confidence intervals for the corresponding vector of multinomial proportions. Recent methods by Fay and Proschan (2015) could be used to combine confidence intervals constructed for each multinomial sample in order to obtain confidence intervals for comparisons between groups, e.g., differences or ratios of proportions. Again, this will not result in intervals for those parameters described above. Exact binomial confidence intervals are described to be conservative for small samples size. Still, they would need a Bonferroni-correction to be adjusted for simultaneous inference and the coverage probability of such intervals remains to be investigated.

The methods considered here are conservative for small samples and extreme proportions, and exact confidence interval methods with Bonferroni-correction should be conservative as well. In situations where primary interest is in hypothesis tests alone, various approaches for stepwise corrections of p-values for multiple comparisons may have higher power than the single step approaches in this paper. However, this entails that corresponding simultaneous confidence intervals are difficult to interpret or to construct (e.g., Strassburger and Bretz, 2008; Schmidt and Brannath, 2015). Further methods for p-value adjustment ensuring FWER control have been explicitly customized for application to sparse discrete data: for applications to multivariate binary, two-sample multinomial, dichotomized multivariate data and permutation approaches without distributional assumptions, Westfall and Wolfinger (1997), Westfall and Troendle (2008) and Westfall (2011) show that these methods can lead to substantially increasing power for sparse discrete multivariate data. In our setting, sparse data would arise from comparing multinomial samples that involve many rare categories which are each of inferential interest. Due to the discreteness of multinomial data and the related test statistics, rejection of the null-hypotheses in an exact test would hardly be possible at all (even without multiplicity adjustment), i.e., type-I-errors are rather improbable to occur for tests of such rare categories. In the simultaneous confidence interval methods considered in this paper, this problem is not taken into account: an 'additional' parameter in the set will increase the dimension (M) of the corresponding multivariate distribution by 1, and thus will lead to more strict adjustment for all multiple comparisons in the set (depending on the correlation of the 'added' parameter to those already in the set). Adjusted p-values based on the multivariate normal distribution (Hothorn et al., 2008) that directly correspond to the Wald-type simultaneous confidence intervals described in this paper, will suffer from the same problem. On the contrary, the adjustment of the given individual p-values following Westfall and Wolfinger (1997), Westfall and Troendle (2008) and Westfall (2011) may not change at all, if the test statistic corresponding to the 'added' category happens to be highly discrete (because the marginal rate of the corresponding category is rare). Thus, such methods may provide higher power compared to the interval approaches discussed here, particularly, if there is a substantial number of rare categories involved as separate tests in the family. For the computational details we refer to Westfall and Wolfinger (1997) and Westfall (2011), an application to the developmental toxicity data in Section 4.1 is available in the supplementary material to this paper (see Appendix A).

One clear limitation of the described methods is that many parameters are fitted to the data. Such approaches may over-fit the data in situations where simpler models would be appropriate. For example, in dose–response analysis, linear or log-linear relations to baseline logits are plausible, and regression models for baseline logits are a sparse alternative to estimating separate parameters for each dose group (see, e.g. Agresti, 2013). In the open source software R, such models are computationally available in package `nnet` (Venables and Ripley, 2002), and simultaneous confidence intervals for linear

Please cite this article in press as: Schaarschmidt, F., et al., Simultaneous confidence intervals for comparisons of several multinomial samples. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.09.004

combinations of their parameters can be obtained by the `multcomp` package (Hothorn et al., 2008). These regression models offer the advantage of a simpler interpretation of relatively few regression slopes instead of many comparisons between individual dose levels. Moreover, the effects of extreme events in single cells of the contingency table should be less extreme than in the methods considered above. As an example, compare the analysis in Section 4.1 with the regression model used in Agresti (1990). When there are several ordinal categories, cumulative logit models or related approaches can be more appropriate (see, e.g. Ryu, 2009; Agresti, 2013).

Furthermore, the methods described here as well as the simulation settings apply only to highly controlled laboratory experiments or randomized trials with no further substructures. However, the Wald-type intervals can likewise be applied when baseline logits and the corresponding covariance matrix are estimated from generalized linear model fits. Then, similar inferential procedures can be performed while including covariates, secondary factors of interest or stratification. Moreover, experiments or studies will often involve replicated biological units per treatment group. For example, several animals, litters, or cultures per treatment group lead to grouped observations in toxicological studies, clustered observations may occur in clinical trials or exposure studies. If variation between these units is larger than expected under multinomial distribution (over-dispersion): all methods considered here will have (severely) too narrow confidence intervals, that is too low coverage probability, because they do not account for such over-dispersion.

Acknowledgments

Q4 We thank three anonymous referees for their constructive comments on earlier versions of the manuscript.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2016.09.004.

References

Agresti, A., 1990. Categorical Data Analysis, first ed. John Wiley & Sons, New York.
Agresti, A., 2013. Categorical Data Analysis, third ed. John Wiley & Sons, Inc., Hoboken, New Jersey.
Besag, J., Green, P., Higdon, D., Mengersen, K., 1995. Bayesian computation and stochastic systems. Statist. Sci. 10, 3–41.
Bretz, F., Genz, A., Hothorn, L., 2001. On the numerical availability of multiple comparison procedures. Biom. J. 43, 645–656.
Casella, G., Berger, R., 2002. Statistical Inference, second ed. Duxbury, Pacific Grove, CA, USA.
Chafai, D., Concordet, D., 2009. Confidence regions for the multinomial parameter with small sample size. J. Amer. Statist. Assoc. 104, 1071–1079.
Chan, I., Zhang, Z., 1999. Test-based exact confidence intervals for the difference of two binomial proportions. Biometrics 55, 1202–1209.
Fay, M., Proschan, M.A., , a.B.E., 2015. Combining one-sample confidence procedures for inference in the two-sample case. Biometrics 71, 146–156.
Genz, A., Bretz, F., 2009. Computation of Multivariate Normal and t Probabilities. In: Lecture Notes in Statistics, vol. 195. Springer-Verlag, Heidelberg.
Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T., 2015. mvtnorm: Multivariate Normal and t Distributions. R package version 1.0-3. URL http://CRAN.R-project.org/package=mvtnorm.
Glaz, J., Sison, C., 1999. Simultaneous confidence intervals for multinomial proportions. J. Statist. Plann. Inference 82, 251–262.
Gold, R., 1963. Test auxiliary to χ^2 in a markov chain. Ann. Math. Statist. 34, 56–74.
Goodman, L., 1964. Simultaneous confidence limits for cross-product ratios in contingency tables. J. R. Stat. Soc. Ser. B Stat. Methodol. 26, 86–102.
Hayter, A., 2014. Recursive formulas for multinomial probabilities with applications. Comput. Statist. 29, 1207–1219.
Hothorn, T., Bretz, F., Westfall, P., 2008. Simultaneous inference in general parametric models. Biom. J. 50, 346–363.
Hothorn, L., Gerhard, D., Pras-Raves, M., 2009. Statistical evaluation of the differential blood count in toxicological studies. Tech. rep.. Institute of Biostatistics, Hannover.
Hou, C.-D., Chiang, J., Tai, J., 2003. A family of simultaneous confidence intervals for multinomial proportions. Comput. Statist. Data Anal. 43, 29–45.
Mandel, M., Betensky, R., 2008. Simultaneous confidence intervals based on the percentile bootstrap approach. Comput. Statist. Data Anal. 52, 2158–2165.
Martin, A., Quinn, K., Park, J., 2011. Mcmcpack: Markov chain monte carlo in r. J. Stat. Softw. 42 (9), 1–21. URL http://www.jstatsoft.org/v42/i09/.
McCullagh, P., Nelder, J., 1989. Generalized Linear Models. Chapman & Hall/CRC.
Piegorsch, W., Richwine, K., 2001. Large-sample pairwise comparisons among multinomial proportions with any application to analysis of mutant spectra. J. Agric. Biol. Environ. Stat. 6, 305–325.
Plackett, R., 1962. A note on interactions in contingency tables. J. R. Stat. Soc. Ser. B Stat. Methodol. 24, 162–166.
Reiczigel, J., Abonyi-Toth, Z., Singer, J., 2008. An exact confidence set for two binomial proportions and exact unconditional confidence intervals for the difference and ratio. Comput. Statist. Data Anal. 52, 5046–5053.
Ryu, E., 2009. Simultaneous confidence intervals using ordinal effect measures for ordered categorical outcomes. Stat. Med. 28, 3179–3188.
Schaarschmidt, F., 2013. Simultaneous confidence intervals for multiple comparisons among expected values of log-normal variables. Comput. Statist. Data Anal. 58, 265–275.
Q5
Q6 Schaarschmidt, F., Djira, G., 2016. Simultaneous confidence intervals for ratios of fixed effect parameters in linear mixed models. Comm. Statist. Simulation Comput. (in press).
Schaarschmidt, F., Gerhard, D., Sill, M., 2016. MCPAN: Multiple Comparisons Using Normal Approximation. R package version 1.1-20. URL http://CRAN.R-project.org/package=MCPAN.
Schmidt, S., Brannath, W., 2015. Informative simultaneous confidence intervals for the fallback procedure. Biom. J. 57, 712–719.
Strassburger, K., Bretz, F., 2008. Compatible simultaneous lower confidence bounds for the holm procedure and other bonferroni-based closed tests. Stat. Med. 27, 4914–4927.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, fourth ed. Springer.
Wang, H., 2000. Exact confidence coefficients of simultaneous confidence intervals for multinomial proportions. J. Multivariate Anal. 99, 896–911.
Wang, W., Shan, G., 2015. Exact confidence intervals for the relative risk and the odds ratio. Biometrics 71, 985–995.
Westfall, P.H., 2011. Improving power by dichotomizing (even under normality). Stat. Biopharm. Res. 3, 353–362.
Westfall, P., Troendle, J., 2008. Multiple testing with minimal assumptions. Biom. J. 50, 745–755.
Westfall, P.H., Wolfinger, R., 1997. Multiple tests with discrete distributions. Amer. Stat. 51, 3–8.